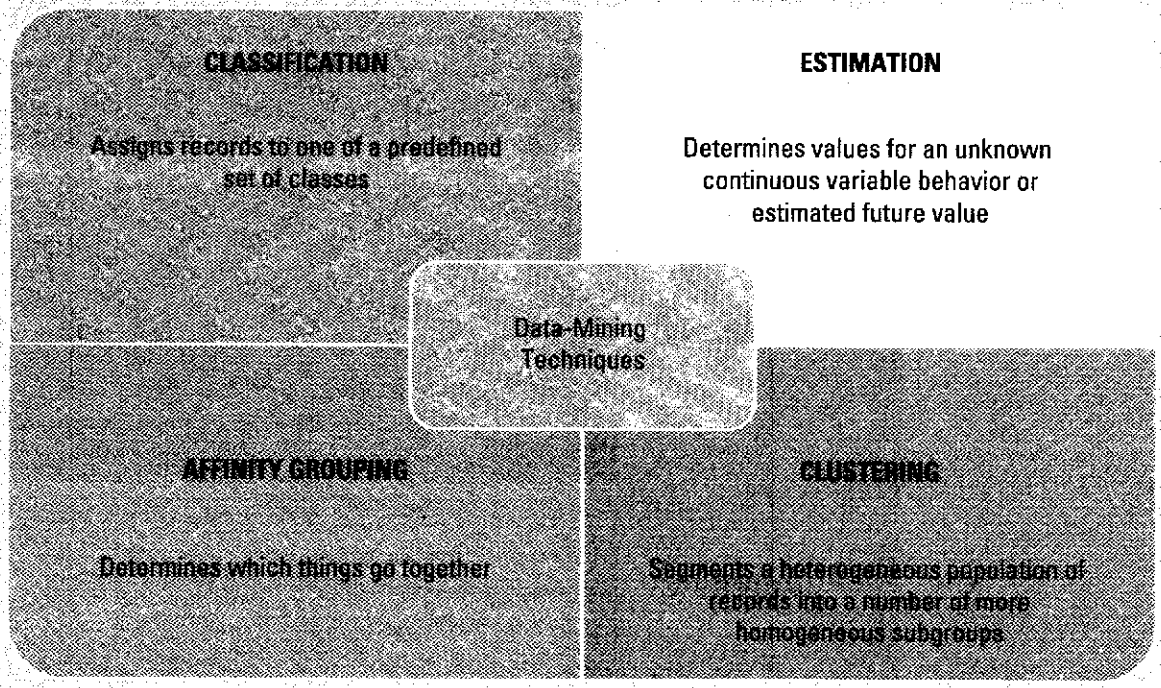
| DATA MINING ANALYSIS METHODS | |
|---|---|
| Prediction | A statement about what will happen or might happen in the future, for example, predicting future sales or employee turnover. |
| Optimization | A statistical process that finds the way to make a design, system, or decision as effective as possible, for example, finding the values of controllable variables that determine maximal productivity or minimal waste. |
| Forecasting | ***Time-series information*** is time-stamped information collected at a particular frequency. Formally defined, forecasts are predictions based on time-series information. Examples of time-series information include web visits per hour, sales per month, and calls per day. Forecasting data-mining tools allow users to manipulate the time series for forecasting activities. |
| Regression | A statistical process for estimating the relationships among variables. It includes many techniques for modeling and analyzing several variables when the focus is on the relationship between a dependent variable and one or more independent variables. |

**FIGURE 6.26**
Data Mining Analysis Methods

Data mining enables these companies to determine relationships among such internal factors as price, product positioning, or staff skills, and external factors such as economic indicators, competition, and customer demographics. In addition, it enables companies to determine the impact on sales, customer satisfaction, and corporate profits and to drill down into summary information to view detailed transactional data. With data mining, a retailer could use point-of-sale records of customer purchases to send targeted promotions based on an individual's purchase history. By mining demographic data from comment or warranty cards, the retailer could develop products and promotions to appeal to specific customer segments.

Netflix uses data mining to analyze each customer's film-viewing habits to provide recommendations for other customers with Cinematch, its movie recommendation system. Using Cinematch, Netflix can present customers with a number of additional movies they might want to watch based on the customer's current preferences. Netflix's innovative use of data mining provides its competitive advantage in the movie rental industry. Data mining uses specialized technologies and functionalities such as query tools, reporting tools, multidimensional analysis tools, statistical tools, and intelligent agents to uncover patterns displayed in Figure 6.27.

**FIGURE 6.27**
Data-Mining Techniques



## Big Data Analytics

***Structured data*** has a defined length, type, and format and includes numbers, dates, or strings such as Customer Address. Structured data is typically stored in a traditional system such as a relational database or spreadsheet and accounts for about 20 percent of the data that surrounds us. The sources of structured data include:

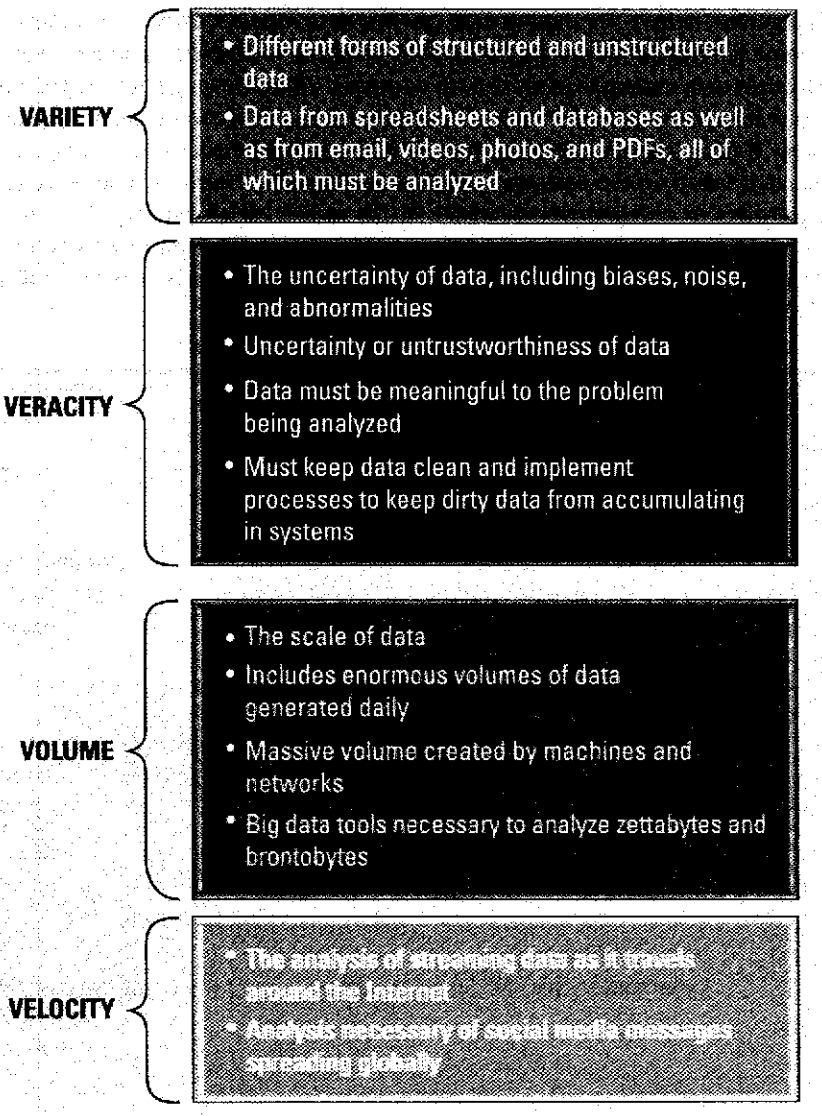
- ***Machine-generated data***, created by a machine without human intervention. Machine-generated structured data includes sensor data, point-of-sale data, and web log (blog) data.
- ***Human-generated data*** is data that humans, in interaction with computers, generate. Human-generated structured data includes input data, click-stream data, or gaming data.

***Unstructured data*** is not defined, does not follow a specified format, and is typically free-form text such as emails, Twitter tweets, and text messages. Unstructured data accounts for about 80 percent of the data that surrounds us. The sources of unstructured data include:

- Machine-generated unstructured data: satellite images, scientific atmosphere data, and radar data.
- Human-generated unstructured data: text messages, social media data, and emails.

***Big data*** is a collection of large, complex data sets, including structured and unstructured data, which cannot be analyzed using traditional database methods and tools. The four common characteristics of big data are detailed in Figure 6.28. Big data requires sophisticated

**FIGURE 6.28**
Four Common Characteristics of Big Data



tools to analyze all the unstructured information from millions of customers, devices, and machine interactions. Big data are analyzed for marketing trends in business as well as in the fields of manufacturing, medicine, and science.

***Distributed computing*** processes and manages algorithms across many machines in a computing environment. Big data tools use distributed computing to store and analyze data across databases stored around the globe. Traditional analytical tools focus on basic business intelligence, including querying and reporting of historical data against a relational database. Traditional data-mining tools focus on history and explain where the organization has been. ***Advanced analytics*** focuses on forecasting future trends and producing insights using sophisticated quantitative methods, including statistics, descriptive and predictive data mining, simulation, and optimization. Advanced analytics uses data patterns to make forward-looking predictions to explain to the organization where it is headed. A ***data scientist*** extracts knowledge from data by performing statistical analysis, data mining, and advanced analytics on big data to identify trends, market changes, and other relevant information. Figure 6.29 displays the techniques a data scientist will use to perform big data advanced analytics.

## Data Visualization

Traditional bar graphs and pie charts are boring and at best confusing and at worst misleading. As databases and graphics collide more and more, people are creating infographics (information graphics), which display information graphically so it can be easily understood. ***Infographics*** present the results of data analysis, displaying the patterns, relationships, and trends in a graphical format. Infographics are exciting and quickly convey a story users can understand without having to analyze numbers, tables, and boring charts. Great data visualizations provide insights into something new about the underlying patterns and relationships. Just think of the periodic table of elements and imagine if you had to look at an Excel spreadsheet showing each element and the associated attributes in a table format. This would be not only difficult to understand but easy to misinterpret. By placing the elements in the visual periodic

**FIGURE 6.29**
Big Data Advanced Analytical Techniques

| BIG DATA ADVANCED ANALYTICAL TECHNIQUES | |
|---|---|
| Association detection | Reveals the relationship between variables along with the nature and frequency of the relationships. Many people refer to association detection algorithms as association rule generators because they create rules to determine the likelihood of events occurring together at a particular time or following each other in a logical progression. Percentages usually reflect the patterns of these events. For example, "55 percent of the time, events A and B occurred together," or "80 percent of the time that items A and B occurred together, they were followed by item C within three days." |
| Cluster analysis | A technique used to divide information sets into mutually exclusive groups such that the members of each group are as close as possible to one another and the different groups are as far apart as possible. Cluster analysis segments customer information to help organizations identify customers with similar behavioral traits, such as clusters of best customers or onetime customers. Cluster analysis also can uncover naturally occurring patterns in information.<br>A great example of using cluster analysis in business is to create target-marketing strategies based on zip codes. Evaluating customer segments by zip code allows a business to assign a level of importance to each segment. Zip codes offer valuable insight into such things as income levels, demographics, lifestyles, and spending habits. With target marketing, a business can decrease its costs while increasing the success rate of the marketing campaign. |
| Market basket analysis | Analyzes such items as websites and checkout scanner information to detect customers' buying behavior and predict future behavior by identifying affinities among customers' choices of products and services.<br>Market basket analysis is frequently used to develop marketing campaigns for cross-selling products and services (especially in banking, insurance, and finance) and for inventory control, shelf product placement, and other retail and marketing applications. |
| Social media analytics | Analyzes text flowing across the Internet, including unstructured text from blogs and messages. |
| Speech analytics | The process of analyzing recorded calls to gather information; brings structure to customer interactions and exposes information buried in customer contact center interactions with an enterprise.<br>Speech analytics is heavily used in the customer service department to help improve processes by identifying angry customers and routing them to the appropriate customer service representative. |
| Text analytics | Analyzes unstructured data to find trends and patterns in words and sentences. Text mining a firm's customer support email might identify which customer service representative is best able to handle the question, allowing the system to forward it to the right person. |
| Web analytics | Analyzes unstructured data associated with websites to identify consumer behavior and website navigation. |

# BUSINESS DRIVEN GLOBALIZATION

## Integrity Information Inc.

Congratulations! You have just been hired as a consultant for Integrity Information Inc., a start-up business intelligence consulting company. Your first job is to help work with the sales department in securing a new client, The Warehouse. The Warehouse has been operating in the United States for more than a decade, and its primary business is to sell wholesale low-cost products. The Warehouse is interested in hiring Integrity Information Inc. to clean up the data that are stored in its U.S. database. To determine how good your work is, the client would like your analysis of the following spreadsheet. The Warehouse is also interested in expanding globally and wants to purchase several independent wholesale stores located in Australia, Thailand, China, Japan, and the United Kingdom. Before the company moves forward with the venture, it wants to understand what types of data issues it might encounter as it begins to transfer data from each global entity to the data warehouse. Please create a list detailing the potential issues The Warehouse can anticipate encountering as it consolidates the global databases into a single data warehouse.[8]

| CUST ID | First Name | Last Name | Address | City | State | Zip | Phone | Last Order Date |
|---|---|---|---|---|---|---|---|---|
| 233620 | Christopher | Lee | 12421 W Olympic Blvd | Los Angeles | CA | 75080-1100 | (972)680-7848 | 4/18/2014 |
| 233621 | Bruce | Brandwen | 268 W 44th St | New York | PA | 10036-3906 | (212)471-6077 | 5/3/2014 |
| 233622 | Glr | Johnson | 4100 E Dry Creek Rd | Littleton | CO | 80122-3729 | (303)712-5461 | 5/6/2014 |
| 233623 | Dave | Owens | 466 Commerce Rd | Staunton | VA | 24401-4432 | (540)851-0362 | 3/19/2014 |
| 233624 | John | Coulbourn | 124 Action St | Maynard | MA | 1754 | (978)987-0100 | 4/24/2014 |
| 233629 | Dan | Gagliardo | 2875 Union Rd | Cheektowaga | NY | 14227-1461 | (716)558-8191 | 5/4/2014 |
| 23362 | Damanceee | Allen | 1633 Broadway | New York | NY | 10019-6708 | (212)708-1576 | |
| 233630 | Michael | Peretz | 235 E 45th St | New York | NY | 10017-3305 | (212)210-1340 | 4/30/2014 |
| 233631 | Jody | Veeder | 440 Science Dr | Madison | WI | 53711-1064 | (608)238-9690 X227 | 3/27/2014 |
| 233632 | Michael | Kehrer | 3015 SSE Loop 323 | Tyler | TX | 75701 | (903)579-3229 | 4/28/2014 |
| 233633 | Erin | Yoon | 3500 Carillon Pt | Kirkland | WA | 98033-7354 | (425)897-7221 | 3/25/2014 |
| 233634 | Madeline | Shefferly | 4100 E Dry Creek Rd | Littleton | CO | 80122-3729 | (303)486-3949 | 3/33/2014 |
| 233635 | Steven | Conduit | 1332 Enterprise Dr | West Chester | PA | 19380-5970 | (610)692-5900 | 4/27/2014 |
| 233636 | Joseph | Kovach | 1332 Enterprise Dr | West Chester | PA | 19380-5970 | (610)692-5900 | 4/28/2014 |
| 233637 | Richard | Jordan | 1700 N | Philadelphia | PA | 19131-4728 | (215)581-6770 | 3/19/2014 |
| 233638 | Scott | Mikolajczyk | 1655 Crofton Blvd | Crofton | MD | 21114-1387 | (410)729-8155 | 4/28/2014 |
| 233639 | Susan | Shragg | 1875 Century Park E | Los Angeles | CA | 90067-2501 | (310)785-0511 | 4/29/2014 |
| 233640 | Rob | Ponto | 29777 Telegraph Rd | Southfield | MI | 48034-1303 | (810)204-4724 | 5/5/2014 |
| 233642 | Lauren | Butler | 1211 Avenue Of The Americas | New York | NY | 10036-8701 | (212)852-7494 | 4/22/2014 |
| 233643 | Christopher | Lee | 12421 W Olympic Blvd | Los Angeles | CA | 90064-1022 | (310)689-2577 | 3/25/2014 |
| 233644 | Michelle | Decker | 6922 Hollywood Blvd | Hollywood | CA | 90028-6117 | (323)817-4655 | 5/8/2014 |
| 233647 | Natalia | Galeano | 1211 Avenue Of The Americas | New York | NY | 10036-8701 | (646)728-6911 | 4/23/2014 |
| 233648 | Bobbie | Orchard | 4201 Congress St | Charlotte | NC | 28209-4617 | (704)557-2444 | 5/11/2014 |
| 233650 | Ben | Konfino | 1111 Stewart Ave | Bethpage | NY | 11714-3533 | (516)803-1406 | 3/19/2014 |
| 233651 | Lenee | Santana | 1050 Techwood Dr NW | Atlanta | GA | 30318-KKRR | (404)885-2000 | 3/22/2014 |
| 233652 | Lauren | Monks | 7700 Wisconsin Ave | Bethesda | MD | 20814-3578 | (301)771-4772 | 3/19/2005 |
| 233653 | Mark | Woolley | 10950 Washington Blvd | Culver City | CA | 90232-4026 | (310)202-2900 | 4/20/2014 |

table, you quickly grasp how the elements relate and the associated hierarchy. Infographics perform the same function for business data as the periodic table does for chemical elements.

***Analysis paralysis*** occurs when the user goes into an emotional state of over-analysis (or over-thinking) a situation so that a decision or action is never taken, in effect paralyzing the outcome. In the time of big data, analysis paralysis is a growing problem. One solution is to use data visualizations to help people make decisions faster. ***Data visualization*** describes technologies that allow users to see or visualize data to transform information into a business perspective. Data visualization is a powerful way to simplify complex data sets by placing data in a format that is easily grasped and understood far quicker than the raw data alone. ***Data visualization tools*** move beyond Excel graphs and charts into sophisticated analysis techniques such as controls, instruments, maps, time-series graphs, and more. Data visualization tools can help uncover correlations and trends in data that would otherwise go unrecognized. ***Business intelligence dashboards*** track corporate metrics such as critical success factors and key performance indicators and include advanced capabilities such as interactive controls, allowing users to manipulate data for analysis. The majority of business intelligence software vendors offer a number of data visualization tools and business intelligence dashboards. A ***data artist*** is a business analytics specialist who uses visual tools to help people understand complex data.

Big data is one of the most promising technology trends occurring today. Of course, notable companies such as Facebook, Google, and Netflix are gaining the most business insights from big data currently, but many smaller markets are entering the scene, including retail, insurance, and health care. Over the next decade, as big data starts to improve your everyday life by providing insights into your social relationships, habits, and careers, you can expect to see the need for data scientists and data artists dramatically increase.

## LEARNING OUTCOME REVIEW

### Learning Outcome 6.1: Explain the four primary traits that determine the value of information.

Information is data converted into a meaningful and useful context. Information can tell an organization how its current operations are performing and help it estimate and strategize about how future operations might perform. It is important to understand the different levels, formats, and granularities of information along with the four primary traits that help determine the value of information, which include (1) information type: transactional and analytical; (2) information timeliness; (3) information quality; and (4) information governance.

### Learning Outcome 6.2: Describe a database, a database management system, and the relational database model.

A database maintains information about various types of objects (inventory), events (transactions), people (employees), and places (warehouses). A database management system (DBMS) creates, reads, updates, and deletes data in a database while controlling access and security. A DBMS provides methodologies for creating, updating, storing, and retrieving data in a database. In addition, a DBMS provides facilities for controlling data access and security, allowing data sharing and enforcing data integrity. The relational database model allows users to create, read, update, and delete data in a relational database.

### Learning Outcome 6.3: Identify the business advantages of a relational database.

Many business managers are familiar with Excel and other spreadsheet programs they can use to store business data. Although spreadsheets are excellent for supporting some data analysis, they offer limited functionality in terms of security, accessibility, and flexibility and can rarely scale to support business growth. From a business perspective, relational databases offer many advantages over using a text document or a spreadsheet, including increased flexibility, increased scalability and performance, reduced information redundancy, increased information integrity (quality), and increased information security.

### Learning Outcome 6.4: Explain the business benefits of a data-driven website.

A data-driven website is an interactive website kept constantly updated and relevant to the needs of its customers using a database. Data-driven capabilities are especially useful when the website offers a great deal of information, products, or services because visitors are frequently annoyed if they are buried under an avalanche of information when searching a website. Many companies use the web to make some of the information in their internal databases available to customers and business partners.

### Learning Outcome 6.5: Identify the advantages of using business intelligence to support managerial decision making.

Many organizations today find it next to impossible to understand their own strengths and weaknesses, let alone their biggest competitors, due to enormous volumes of organizational data being inaccessible to all but the MIS department. Organization data include far more than simple structured data elements in a database; the set of data also includes unstructured data such as voice mail, customer phone calls, text messages, video clips, along with numerous new forms of data, such as tweets from Twitter. Managers today find themselves in the position of being data rich and information poor, and they need to implement business intelligence systems to solve this challenge.

### Learning Outcome 6.6: Define data warehousing and data marts and explain how they support business decisions.

A data warehouse is a logical collection of information, gathered from many different operational databases, that supports business analysis and decision making. The primary value of a data warehouse is to combine information, more specifically, strategic information, throughout an organization into a single repository in such a way that the people who need that information can make decisions and undertake business analysis.

### Learning Outcome 6.7: Describe the three organizational methods for analyzing big data.

Data mining, big data analytics, and data visualization are the three methods organizations are using to dissect, analyze, and understand organizational data. Data mining is the process of analyzing data to extract information not offered by the raw data alone. Data mining can also begin at a summary information level (coarse granularity) and progress through increasing levels of detail (drilling down), or the reverse (drilling up). Big data is a collection of large, complex data sets, including structured and unstructured data, which cannot be analyzed using traditional database methods and tools. Data visualization describes technologies that allow users to see or visualize data to transform information into a business perspective.

## OPENING CASE QUESTIONS

1. **Knowledge:** List the reasons a business would want to display information in a graphic or visual format.
2. **Comprehension:** Describe how a business could use a business intelligence digital dashboard to gain an understanding of how the business is operating.
3. **Application:** Explain how a marketing department could use data visualization tools to help with the release of a new product.
4. **Analysis:** Categorize the five common characteristics of high-quality information and rank them in order of importance for Hotels.com.
5. **Synthesis:** Develop a list of some possible entities and attributes located in the Hotels.com database.
6. **Evaluate:** Assess how Hotels.com is using BI to identify trends and change associated business processes.

## KEY TERMS

Advanced analytics, 241
Analysis paralysis, 243
Attribute, 222
Big data, 240
Business-critical integrity constraint, 226
Business rule, 226
Business intelligence dashboard, 243
Content creator, 227
Content editor, 227
Data dictionary, 221
Data element (or data field), 221
Data governance, 220
Data mart, 233
Data mining, 238
Data model, 221
Data quality audit, 237
Data visualization, 243
Data visualization tools, 243
Data warehouse, 231
Database, 220
Database management system (DBMS), 220
Data-driven website, 227
Data-mining tool, 238
Data artist, 243
Data scientist, 241
Dirty data, 234
Distributed computing, 241
Dynamic catalog, 227
Dynamic information, 227
Entity, 222
Extraction, transformation, and loading (ETL), 233
Foreign key, 223
Human-generated data, 240
Infographic (or information graphic), 241
Information cleansing or scrubbing, 235
Information cube, 234
Information granularity, 215
Information inconsistency, 217
Information integrity, 226
Information integrity issues, 217
Information redundancy, 226
Integrity constraint, 226
Logical view of information, 225
Machine-generated data, 240
Master data management (MDM), 220
Metadata, 221
Physical view of information, 225
Primary key, 222
Query-by-example (QBE) tool, 221
Real-time information, 217
Real-time system, 217
Record, 222
Relational database management system, 222
Relational database model, 221
Relational integrity constraint, 226
Static information, 227
Structured data, 240
Structured query language (SQL), 221
Time-series information, 239
Unstructured data, 240

## REVIEW QUESTIONS

1. How does a database turn data elements into information?
2. Why does a business need to be concerned with the quality of its data?
3. How can data governance help protect a business from hackers?
4. Why would a company care about the timeliness of its data?
5. What are the five characteristics common to high-quality information?
6. What is data governance and its importance to a company?
7. What are the four primary traits that help determine the value of information?
8. What is the difference between an entity and an attribute?
9. What are the advantages of a relational database?
10. What are the advantages of a data-driven website?
11. What is a data warehouse and why would a business want to implement one?
12. Why would you need to use multidimensional analysis?
13. What is the purpose of information cleansing (or scrubbing)?
14. Why would a department want a data mart instead of just accessing the entire data warehouse?
15. Why would a business be data rich but information poor?

## CLOSING CASE ONE

# Data Visualization: Stories for the Information Age

At the intersection of art and algorithm, data visualization schematically abstracts information to bring about a deeper understanding of the data, wrapping it in an element of awe. Although the practice of visually representing information is arguably the foundation of all design, a newfound fascination with data visualization has been emerging. After *The New York Times* and *The Guardian* recently opened their online archives to the public, artists rushed to dissect nearly two centuries' worth of information, elevating this art form to new prominence.

For artists and designers, data visualization is a new frontier of self-expression, powered by the proliferation of information and the evolution of available tools. For enterprise, it is a platform for displaying products and services in the context of the cultural interaction that surrounds them, reflecting consumers' increasing demand for corporate transparency.

"Looking at something ordinary in a new way makes it extraordinary," says Aaron Koblin, one of the more recent pioneers of the discipline. As technology lead of Google's Creative Labs in San Francisco, he spearheaded the search giant's Chrome Experiments series designed to show off the speed and reliability of the Chrome browser.

## Forget Pie Charts and Bar Graphs

Data visualization has nothing to do with pie charts and bar graphs. And it's only marginally related to infographics, information design that tends to be about objectivity and clarification. Such representations simply offer another iteration of the data—restating it visually and making it easier to digest. Data visualization, on the other hand, is an interpretation, a different way to look at and think about data that often exposes complex patterns or correlations.

Data visualization is a way to make sense of the ever-increasing stream of information with which we're bombarded and provides a creative antidote to the analysis paralysis that can result from the burden of processing such a large volume of information. "It's not about clarifying data," says Koblin. "It's about contextualizing it."

Today algorithmically inspired artists are reimagining the art-science continuum through work that frames the left-brain analysis of data in a right-brain creative story. Some use data visualization as a bridge between alienating information and its emotional impact—see Chris Jordan's portraits of global mass culture. Others take a more technological angle and focus on cultural utility—the Zoetrope project offers a temporal and historical visualization of the ephemeral web. Still others are pure artistic indulgence—like Koblin's own Flight Patterns project, a visualization of air traffic over North America.

## How Business Can Benefit

There are real implications for business here. Most cell phone providers, for instance, offer a statement of a user's monthly activity. Most often it's an overwhelming table of various numerical measures of how much you talked, when, with whom, and how much it cost. A visual representation of this data might help certain patterns emerge, revealing calling habits and perhaps helping users save money.

Companies can also use data visualization to gain new insight into consumer behavior. By observing and understanding what people do with the data—what they find useful and what they dismiss as worthless—executives can make the valuable distinction between what consumers say versus what they do. Even now, this can be a tricky call to make from behind the two-way mirror of a traditional qualitative research setting.

It's essential to understand the importance of creative vision along with the technical mastery of software. Data visualization isn't about using all the data available, but about deciding which patterns and elements to focus on, building a narrative, and telling the story of the raw data in a different, compelling way.

Ultimately, data visualization is more than complex software or the prettying up of spreadsheets. It's not innovation for the sake of innovation. It's about the most ancient of social rituals: storytelling. It's about telling the story locked in the data differently, more engagingly, in a way that draws us in, makes our eyes open a little wider and our jaw drop ever so slightly. And as we process it, it can sometimes change our perspective altogether.[9]

### Questions

1. Identify the effects poor information might have on a data visualization project.
2. How does data visualization use database technologies?
3. How could a business use data visualization to identify new trends?
4. What is the correlation between data mining and data visualization?
5. Is data visualization a form of business intelligence? Why or why not?
6. What security issues are associated with data visualization?
7. What might happen to a data visualization project if it failed to cleanse or scrub its data?

## CLOSING CASE TWO

### Zillow

Zillow.com is an online, web-based real estate site helping homeowners, buyers, sellers, renters, real estate agents, mortgage professionals, property owners, and property managers find and share information about real estate and mortgages. Zillow allows users to access, anonymously and free of charge, the kinds of tools and information previously reserved for real estate professionals. Zillow's databases cover more than 90 million homes, which represents 95 percent of the homes in the United States. Adding to the sheer size of its databases, Zillow recalculates home valuations for each property every day, so it can provide historical graphs on home valuations over time. In some areas, Zillow is able to display 10 years of valuation history, a value-added benefit for many of its customers. This collection of data represents an operational data warehouse for anyone visiting the website.

As soon as Zillow launched its website, it immediately generated a massive amount of traffic. As the company expanded its services, the founders knew the key to its success would be the site's ability to process and manage massive amounts of dataquickly, in real time. The company identified a need for accessible, scalable, reliable, secure databases that would enable it to continue to increase the capacity of its infrastructure indefinitely without sacrificing performance. Zillow's traffic continues to grow despite the weakened real estate market; the company is experiencing annual traffic growth of 30 percent, and about a third of all U.S. mortgage professionals visit the site in a given month.

### Data Mining and Business Intelligence

Zestimate values on Zillow use data-mining features for spotting trends across property valuations. Data mining also allows the company to see how accurate Zestimate values are over time. Zillow has also built the industry's first search by monthly payment, allowing users to find homes that are for sale and rent based on a monthly payment they can afford. Along with the monthly payment search, users can also enter search criteria such as the number of bedrooms or bathrooms.

Zillow also launched a new service aimed at changing the way Americans shop for mortgages. Borrowers can use Zillow's new Mortgage Marketplace to get custom loan quotes from lenders without having to give their names, addresses, phone numbers, or Social Security numbers, or field unwanted telephone calls from brokers competing for their business. Borrowers reveal their identities only after contacting the lender of their choice. The company is entering a field of established mortgage sites such as LendingTree.com and Experian Group's Lowermybills.com, which charge

mortgage companies for borrower information. Zillow, which has an advertising model, says it does not plan to charge for leads.

For mortgage companies, the anonymous leads come free; they can make a bid based on information provided by the borrower, such as salary, assets, credit score, and the type of loan. Lenders can browse borrower requests and see competing quotes from other brokers before making a bid.[10]

### Questions

1. List the reasons Zillow would need to use a database to run its business.
2. Describe how Zillow uses business intelligence to create a unique product for its customers.
3. How could the marketing department at Zillow use a data mart to help with the release of a new product launch?
4. Categorize the five common characteristics of high-quality information and rank them in order of importance to Zillow.
5. Develop a list of some possible entities and attributes of Zillow's mortgage database.
6. Assess how Zillow uses a data-driven website to run its business.

## CRITICAL BUSINESS THINKING

### 1. Information—Business Intelligence or a Diversion from the Truth?

President Obama used part of his commencement address at Virginia's Hampton University to criticize the flood of incomplete information or downright incorrect information that flows in the 24-hour news cycle. The president said, "You're coming of age in a 24/7 media environment that bombards us with all kinds of content and exposes us to all kinds of arguments, some of which don't always rank all that high on the truth meter. With iPods and iPads and Xboxes and PlayStations—none of which I know how to work—information becomes a distraction, a diversion, a form of entertainment, rather than a tool of empowerment, rather than the means of emancipation."[11]

Do you agree or disagree with President Obama's statement? Who is responsible for verifying the accuracy of online information? What should happen to companies that post inaccurate information? What should happen to individuals who post inaccurate information? What should you remember when reading or citing sources for online information?

### 2. Illegal Database Access

Goldman Sachs has been hit with a $3 million lawsuit by a company that alleges the brokerage firm stole intellectual property from its database that had market intelligence facts. The U.S. District Court for the Southern District of New York filed the lawsuit in 2010 claiming Goldman Sachs employees used other people's access credentials to log on to Ipreo's proprietary database, dubbed Bigdough. Offered on a subscription basis, Bigdough provides detailed information on more than 80,000 contacts within the financial industry. Ipreo complained to the court that Goldman Sachs employees illegally accessed Bigdough at least 264 times in 2008 and 2009.[12]

Do you agree or disagree with the lawsuit? Should Goldman Sachs be held responsible for rogue employees' behavior? What types of policies should Goldman Sachs implement to ensure that this does not occur again?

### 3. Data Storage

Information is one of the most important assets of any business. Businesses must ensure information accuracy, completeness, consistency, timeliness, and uniqueness. In addition, business must have a reliable backup service. In part thanks to cloud computing, there are many data hosting services on the Internet. These sites offer storage of information that can be accessed from anywhere in the world.

These data hosting services include Hosting (www.hosting.com), Mozy (www.mozy.com), My Docs Online (www.mydocsonline.com), and Box (www.box.net). Visit a few of these sites along with several others you find through research. Which sites are free? Are there limits to how much you can store? If so, what is the limit? What type of information can you store (video, text, photos, etc.)? Can you allow multiple users with different passwords to access your storage area? Are you contractually bound for a certain duration (annual, etc.)? Are different levels of services provided such as personal, enterprise, and work group? Does it make good business sense to store business data on the Internet? What about personal data?

### 4. Gathering Business Intelligence

When considering new business opportunities, you need knowledge about the competition. One of the things many new business owners fail to do is gather business intelligence on their competitors, such as how many there are and what differentiates each of them. You may find there are too many and that they would be tough competition for you. Or, you may find that there are few competitors and the ones who are out there offer very little value.

Generate a new business idea you could launch on the Internet. Research the Internet to find similar business in the area you have chosen. How many sites did you find that are offering the same products or services you are planning to offer? Did you come across any sites from another country that have a unique approach that you did not see on any of the sites in your own country? How would you use this information in pursuing your business idea?

### 5. Free Data!

The U.S. Bureau of Labor Statistics states that its role is as the "principal fact-finding agency for the federal government in the broad field of labor economics and statistics." And the data that the bureau provides via its website are available to anyone, free. This can represent a treasure trove of business intelligence and data mining for those who take advantage of this resource. Visit the website www.bls.gov. What type of information does the site provide? What information do you find most useful? What sort of information concerning employment and wages is available? How is this information categorized? How would this type of information be helpful to a business manager? What type of demographic information is available? How could this benefit a new start-up business?[13]

### 6. Explaining Relational Databases

You have been hired by Vision, a start-up clothing company. Your manager, Holly Henningson, is unfamiliar with databases and their associated business value. Henningson has asked you to create a report detailing the basics of databases. She would also like you to provide a detailed explanation of relational databases along with their associated business advantages.

### 7. Entities and Attributes

Martex Inc. is a manufacturer of athletic equipment, and its primary lines of business include running, tennis, golf, swimming, basketball, and aerobics equipment. Martex currently supplies four primary vendors, including Sam's Sports, Total Effort, The Underline, and Maximum Workout. Martex wants to build a database to help it organize its products. In a group, identify the different types of entities, attributes, keys, and relationships Martex will want to consider when designing its relational database.

### 8. Compiling Information

You are currently working for the Public Transportation Department of Chatfield. The department controls all forms of public transportation, including buses, subways, and trains. Each department has about 300 employees and maintains its own accounting, inventory, purchasing, and human resource systems. Generating reports across departments is a difficult task and usually involves gathering and correlating the information from the many databases. It typically takes about two weeks to generate the quarterly balance sheets and profit and loss statements. Your team has been asked to compile a report recommending what the Public Transportation Department of Chatfield can do to alleviate its information and system issues. Be sure that your report addresses

the various reasons departmental reports are presently difficult to obtain as well as how you plan to solve this problem.[14]

**9. Information Timeliness**

Information timeliness is a major consideration for all organizations. Organizations need to decide the frequency of backups and the frequency of updates to a data warehouse. In a team, describe the timeliness requirements for backups and updates to a data warehouse for each of the following:

- Weather tracking systems
- Car dealership inventories
- Vehicle tire sales forecasts
- Interest rates
- Restaurant inventories
- Grocery store inventories

**10. Improving Information Quality**

HangUps Corporation designs and distributes closet organization structures. The company operates five systems—order entry, sales, inventory management, shipping, and billing. The company has severe information quality issues, including missing, inaccurate, redundant, and incomplete information. The company wants to implement a data warehouse containing information from the five systems to help maintain a single customer view, drive business decisions, and perform multidimensional analysis. Identify how the organization can improve its information quality when it begins designing and building its data warehouse.

## ENTREPRENEURIAL CHALLENGE

### BUILD YOUR OWN BUSINESS

1. Provide an example of your business data that fits each of the five common characteristics of high-quality information. Explain why each characteristic is important to your business data and what might happen if your business data were of low quality. (Be sure to identify your business and the name of your company.)
2. Identify the different entities and their associated attributes that would be found in your potential relational database model for your sales database.
3. Identify the benefits of having a data warehouse for your business. What types of data marts would you want to extract from your data warehouse to help you run your business and make strategic decisions?

## APPLY YOUR KNOWLEDGE BUSINESS PROJECTS

### PROJECT I Mining the Data Warehouse

Alana Smith is a senior buyer for a large wholesaler that sells different types of arts and crafts to greeting card stores such as Hallmark. Smith's latest marketing strategy is to send all of her customers a new line of handmade picture frames from Russia. All of her information supports her decision for the new line. Her analysis predicts that the frames should sell an average of 10 to 15 per store, per day. Smith is excited about the new line and is positive it will be a success.

One month later, Smith learns the frames are selling 50 percent below expectations and averaging between five and eight frames sold daily in each store. She decides to access the company's data warehouse information to determine why sales are below expectations. Identify several dimensions of information that Smith will want to analyze to help her decide what is causing the problems with the picture frame sales.

## PROJECT II Different Dimensions

The focus of data warehousing is to extend the transformation of data into information. Data warehouses offer strategic level, external, integrated, and historical information so businesses can make projections, identify trends, and make key business decisions. The data warehouse collects and stores integrated sets of historical information from multiple operational systems and feeds them to one or more data marts. It may also provide end user access to support enterprisewide views of information.

You are currently working on a marketing team for a large corporation that sells jewelry around the world. Your boss has asked you to look at the following dimensions of data to determine which ones you want in your data mart for performing sales and market analysis (see Figure AYK.1). As a team, categorize the different dimensions, ranking them from 1 to 5, with 1 indicating that the dimension offers the highest value and must be in your data mart and 5 indicating that the dimension offers the lowest value and does not need to be in your data mart.

## PROJECT III Understanding Search

Pretend that you are a search engine. Choose a topic to query. It can be anything such as your favorite book, movie, band, or sports team. Search your topic on Google, pick three or four pages from the results, and print them out. On each printout, find the individual words from your query (such as

**FIGURE AYK.1**
Data Warehouse Data

| Dimension | Value (1–5) | Dimension | Value (1–5) |
|---|---|---|---|
| Product number | | Season | |
| Store location | | Promotion | |
| Customer net worth | | Payment method | |
| Number of sales personnel | | Commission policy | |
| Customer eating habits | | Manufacturer | |
| Store hours | | Traffic report | |
| Salesperson ID | | Customer language | |
| Product style | | Weather | |
| Order date | | Customer gender | |
| Product quantity | | Local tax information | |
| Ship date | | Local cultural demographics | |
| Current interest rate | | Stock market closing | |
| Product cost | | Customer religious affiliation | |
| Customer's political affiliation | | Reason for purchase | |
| Local market analysis | | Employee dress code policy | |
| Order time | | Customer age | |
| Customer spending habits | | Employee vacation policy | |
| Product price | | Employee benefits | |
| Exchange rates | | Current tariff information | |
| Product gross margin | | | |

"Boston Red Sox" or "The Godfather") and use a highlighter to mark each word with color. Do that for each of the documents that you print out. Now tape those documents on a wall, step back a few feet, and review your documents.

If you did not know what the rest of a page said and could judge only by the colored words, which document do you think would be most relevant? Is there anything that would make a document look more relevant? Is it better for the words to be in a large heading or to occur several times in a smaller font? Do you prefer the words to be at the top or the bottom of the page? How often do the words need to appear? Come up with two or three things you would look for to see whether a document matched a query well. This exercise mimics search engine processes and should help you understand why a search engine returns certain results over others.

### PROJECT IV Predicting Netflix

Netflix Inc., the largest online movie rental service, provides more than 12 million subscribers access to more than 100,000 unique DVD titles along with a growing on-demand library in excess of 10,000 choices. Data and information are so important to Netflix that it created The Netflix Prize, an open competition for anyone who could improve the data used in prediction ratings for films (an increase of 10 percent), based on previous ratings. The winner would receive a $1 million prize.

The ability to search, analyze, and comprehend information is vital for any organization's success. It certainly was for Netflix—it was happy to pay anyone $1 million to improve the quality of its information. In a group, explain how Netflix might use databases, data warehouses, and data marts to predict customer movie recommendations. Here are a few characteristics you might want to analyze to get you started:

- Customer demographics
- Movie genre, rating, year, producer, and type
- Actor information
- Internet access
- Location for mail pickup

### PROJECT V The Crunch Factory

The Crunch Factory is one of the fourth-largest gyms operating in Australia, and each gym operates its own system with its own database. Unfortunately, the company failed to develop any data-capturing standards and now faces the challenges associated with low-quality enterprisewide information. For example, one system has a field to capture email addresses, but another system does not. Duplicate customer information among the different systems is another major issue, and the company continually finds itself sending conflicting or competing messages to customers from different gyms. A customer could also have multiple accounts within the company, one representing a membership, another representing additional classes, and yet another for a personal trainer. The Crunch Factory has no way to identify that the different customer accounts are actually for the same customer.

To remain competitive and be able to generate business intelligence, The Crunch Factory has to resolve these challenges. The Crunch Factory has just hired you as its data quality expert. Your first task is to determine how the company can turn its low-quality information into high-quality business intelligence. Create a plan that The Crunch Factory can implement that details the following:

- Challenges associated with low-quality information
- Benefits associated with high-quality information
- Recommendations on how the company can clean up its data

### PROJECT VI Too Much of a Good Thing

The Castle, a premium retailer of clothes and accessories, created an enterprisewide data warehouse so all its employees could access information for decision making. The Castle soon discovered that it

# Designing Databases

APPENDIX C

## LEARNING OUTCOMES

C.1 Identify the relational database model's basic components.

C.2 Explain the importance of documenting entity relationship diagrams.

C.3 Explain the need for an entity-relationship diagram in a database management system.

## INTRODUCTION

LO C.1: Identify the relational database model's basic components.

Businesses rely on databases for accurate, up-to-date information. Without access to mission critical data, most businesses are unable to perform their normal daily functions, much less create queries and reports that help make strategic decisions. For those decisions to be useful, the database must have data that are accurate, complete, consistent, timely, and unique. However, without a good underlying database design, decisions will be inaccurate and inconstant.

A ***database*** maintains information about various types of objects (inventory), events (transactions), people (employees), and places (warehouses). A ***database management system (DBMS)*** creates, reads, updates, and deletes data in a database while controlling access and security. A DBMS provides a way to create, update, delete, store, and retrieve data in the database.

Using a data model offers a method for designing a database correctly that will help in meeting the needs of the users in a DBMS environment.

## THE RELATIONAL DATABASE MODEL

Numerous elements in a business environment need to store data, and those elements are related to one another in a variety of ways. In fact, a database must contain not only the data but also information about the relationships between those data. Designing a database properly is fundamental to establishing a solid foundation on which to base business decisions. This is done by using a ***data model,*** or the logical data structures that detail the relationships among data elements using graphics or pictures. A ***relational database model*** stores information in the form of logically related two-dimensional tables. Tables, or entities as they are formally referred to, will be discussed later.

In developing the relational database model to design a database, an entity-relationship diagram is used. An ***entity-relationship diagram (ERD)*** is a technique for documenting the entities and relationships in a database environment. Before describing the notation used for developing an ERD, it is important to understand what entities and attributes are.

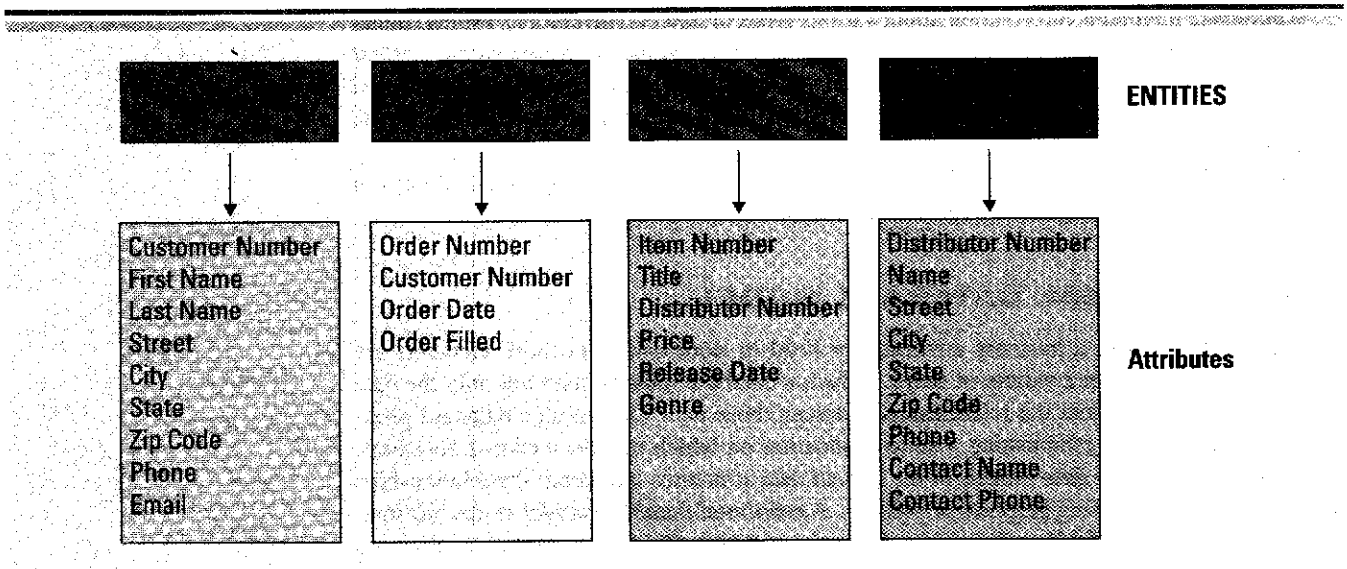
## Entities and Attributes

An ***entity*** stores information about a person, place, thing, transaction, or event. A customer is an entity, as is a product and an appointment. An ***attribute*** is the data elements associated with an entity. For example, consider Mega-Video, a physical and online retail store that sells movie DVDs. The company would need to store information about its customers (especially for online purchases) by creating an entity called *CUSTOMER* that contained many attributes such as *Customer Number, First Name, Last Name, Street, City, State, Zip Code, Phone Number,* and *Email* (refer to Figure C.1).

**Type of Attributes** There are several types of attributes, including:

- **Simple versus composite.** A simple attribute cannot be broken down into a smaller component. For example, a customer's first name and last name are simple. A composite attribute can be divided into smaller components, which represent more basic attributes that have their own meanings. A common example of a composite attribute is *Address* (see Figure C.2). *Address* can be broken down into a number of subparts, such as *Street, City, State, Zip Code.*
- **Single-valued versus multivalued.** When creating a relational database, the attributes in the data model must be single-valued. ***Single-valued attribute*** means having only a single value of each attribute of an entity. A person's age is an example of a single-valued attribute because a person cannot have more than one age. ***Multivalued attribute*** means having the potential to contain more than one value for an attribute. For example, an educational degree of a person is a multivalued attribute because a person can have more than one degree. An entity in a relational database cannot have multivalued attributes; those attributes must be handled by creating another entity to hold them. Therefore, in the example given previously, in designing the database there would be two entities, one called *PERSON* (or something similar) and one called *DEGREE.* If a multivalued attribute has been identified, it typically is a clue that another entity is needed.

**FIGURE C.1**
Entities and Attributes Examples



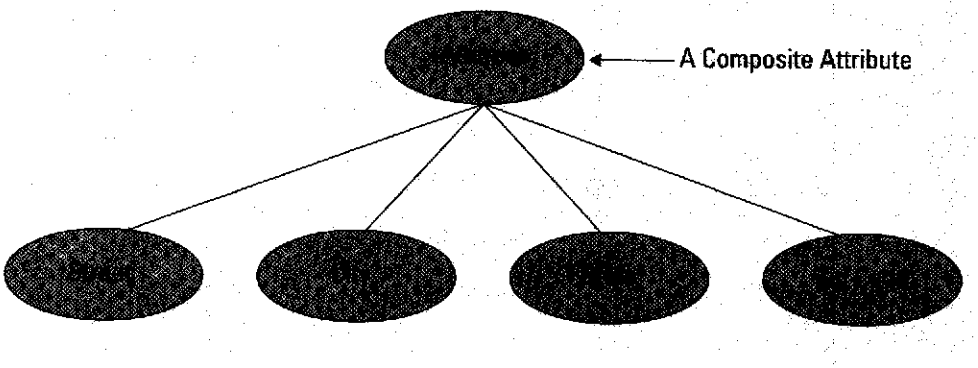



**FIGURE C.2**
Composite Attributes

- **Stored versus derived.** If an attribute can be calculated using the value of another attribute, it is called a derived attribute. The attribute that is used to derive the attribute is called a stored attribute. Derived attributes are not stored in the file but can be derived when needed from the stored attributes. One example of a derived and stored attribute is a person's age. If the database has a stored attribute such as the person's *Date of Birth,* then a derived attribute called *Age* can be created by subtracting the *Current Date* (this is retrieved from the DBMS) from the *Date of Birth* to get the age.
- **Null-valued.** Sometimes an attribute does not have an applicable value for an attribute. For these situations, the null-valued attribute is created. ***Null-valued attribute*** is assigned to an attribute when no other value applies or when a value is unknown. A person who does not have a mobile phone would have null stored for the value for the *Mobile Phone Number* attribute. Null can also be used when the attribute value is unknown, such as *Hair Color.* Every person has a hair color, but the information may be missing.

## Business Rules

The "correct" design for a specific business depends on the business rules; what is correct for one organization may not be correct for another. A ***business rule*** is a statement that defines an aspect of a business. It is intended to convey the behavior and rules of a business. The following statements are examples of possible business rules for Mega-Video:

- A customer can purchase many DVDs.
- DVDs can be purchased by many customers.
- A DVD title can have many copies.

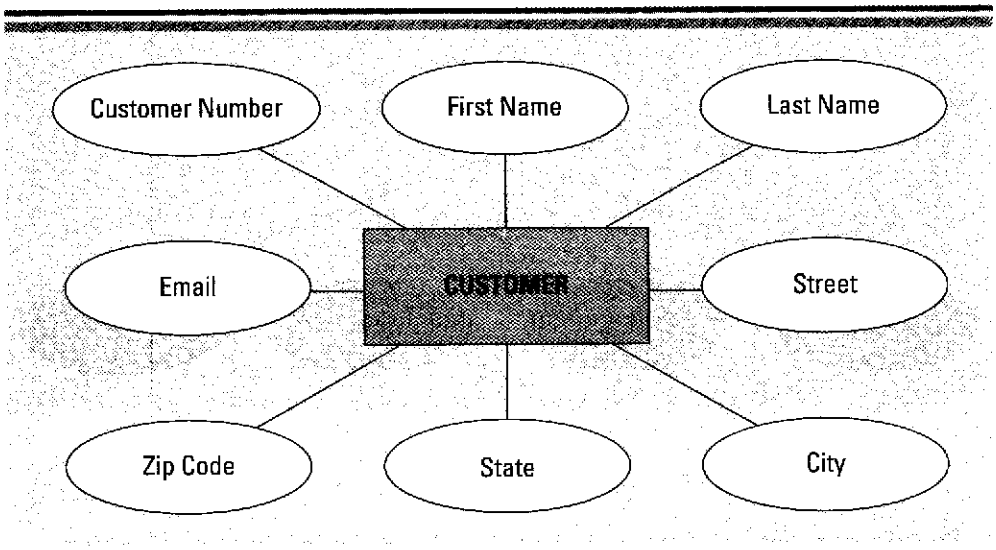
A typical business may have hundreds of business rules. Each business rule will have entities and sometimes even attributes in the statements. For instance, in the preceding example, *CUSTOMER* and *DVD* would be entities according to this business rule. Identifying the business rules will help to create a more accurate and complete database design. In addition, the business rules also assist with identifying relationships between entities. This is very useful in creating ERDs.

# DOCUMENTING ENTITY RELATIONSHIP DIAGRAMS

**LO C.2: Explain the importance of documenting entity relationship diagrams.**

Once entities, attributes, and business rules have been identified, the ERD can be documented. The two most commonly used models of ERD documentation are Chen, named after the originator of entity-relationship modeling, Dr. Peter Chen, and information engineering, which grew out of work by James Martin and Clive Finkelstein. It does not matter

**FIGURE C.3**
Chen Model with Attributes




which is used as long as everyone who is using the diagram understands the notation. For purposes of simplicity, only the Chen model will be described here.

The Chen model notation uses very specific symbols in representing entities and attributes. Rectangles represent entities. Each entity's name appears in the rectangle, is expressed in the singular, and is capitalized, as in *CUSTOMER*. Originally, attributes were not part of the Chen model; however, many database designers have extended it to include the attributes in ovals as illustrated in Figure C.3.

## Basic Entity Relationships

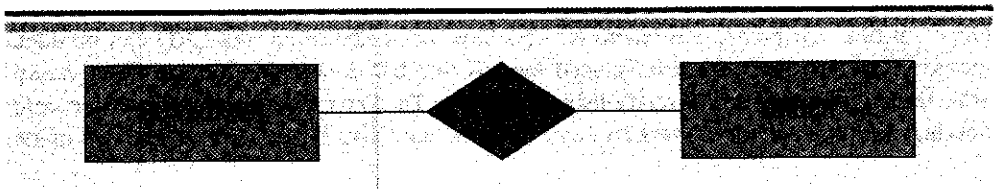
One of the main reasons for creating an ERD is to identify and represent the relationships between entities. If the business rules for Mega-Video state that a customer can order many videos (in this case, an item), then a relationship needs to be created between *CUSTOMER*, *ORDER*, and *ITEM*. This is a purely conceptual representation of what the database will look like and is completely unrelated to the physical storage of the data. Again, what the ERD is doing is creating a model in which to design the database.

The Chen model uses diamonds for relationships and lines to connect relationships between entities. Figure C.4 displays the relationship between a Mega-Video *CUSTOMER* and *ORDER* using this notation. The word within the relationship gives some indication of the meaning of the relationship.

Once the basic entities and their attributes have been defined, the next task is to identify the relationships among those entities. There are three basic types of relationships: (1) one-to-one, (2) one-to-many, and (3) many-to-many.

**One-to-One Relationship** A ***one-to-one relationship (1:1)*** is between two entities in which an instance of one entity can be related to only one instance of a related entity. Consider Mega-Video, which has many stores with several employees and one manager. According to the company's business rules, the manager, who is an employee, can manage only one store. The relationship then becomes 1:1 between *EMPLOYEE* and *STORE*. Using the Chen model notation, as shown in Figure C.5, the relationships between the two instances can then be expressed as "An employee can manage one store and one store has

**FIGURE C.4**
Chen Method with Relationship



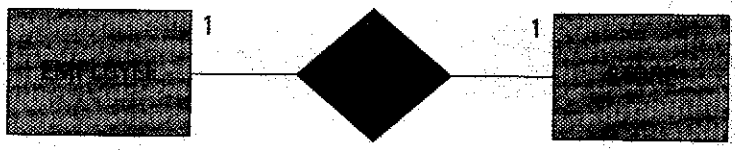



**FIGURE C.5**
A One-to-One Relationship

one manager." The number "1" next to the *EMPLOYEE* and *STORE* entities indicates that only one *EMPLOYEE* manages one *STORE.*

**One-to-Many Relationship** Most relational databases are constructed from one-to-many relationships. A ***one-to-many relationship (1:M)*** is between two entities in which an instance of one entity can be related to many instances of a related entity. For example, Mega-Video receives many *ITEM*(s) from one *DISTRIBUTOR,* and each *DISTRIBUTOR* supplies many *ITEM*(s) as Figure C.6 illustrates. Similarly, a *CUSTOMER* can have many *ORDER*(s), but an *ORDER* has only one *CUSTOMER.* These are both examples of a one-to-many relationship. The "M" next to the *ORDER* entity indicates that a *CUSTOMER* can place one or more *ORDER*(s). That notation is also used with *ITEM,* because an *ORDER* can contain one or more *ITEM*(s).

**Many-to-Many Relationship** Identifying and removing many-to-many relationships will help to create an accurate and consistent database. A ***many-to-many relationship (M:N)*** is between two entities in which an instance of one entity is related to many instances of another and one instance of the other can be related to many instances of the first entity. There is a many-to-many relationship between a Mega-Video *ORDER* and *ITEM* (refer back to Figure C.6). An *ORDER* can contain many *ITEM*(s), and each *ITEM* can be on many *ORDER*(s). The letter "N" next to *ITEM* in Figure C.6 indicates the many-to-many relationship between *ORDER* and *ITEM.*

However, there are problems with many-to-many relationships. First, the relational data model was not designed to handle many-to-many relationships. This means they need to be replaced with a one-to-many relationship to be used in a relational DBMS. Second, many-to-many relationships will create redundancy in the data that are stored. This then has a negative impact on the accuracy and consistency that a database needs. To understand this problem better, consider the relationship between *ITEM* and *ORDER.* There is a many-to-many relationship between the *ORDER* and the *ITEM* because each *ORDER* can contain many *ITEM*(s) and, over time, each *ITEM* will be on many *ORDER*(s). Whenever a *CUSTOMER* places an *ORDER* for an *ITEM,* the number of *ITEM*(s) varies, depending on how many DVDs the *CUSTOMER* is buying. To break the many-to-many relationship, a composite entity is needed.

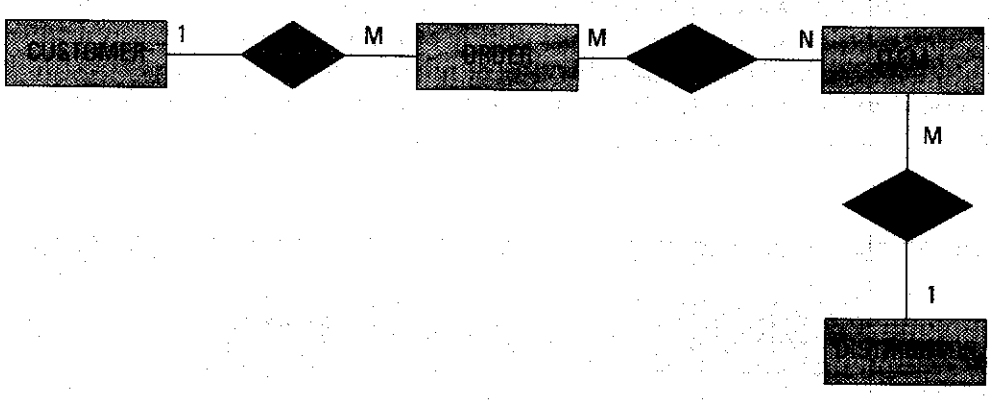



**FIGURE C.6**
A One-to-Many Relationship

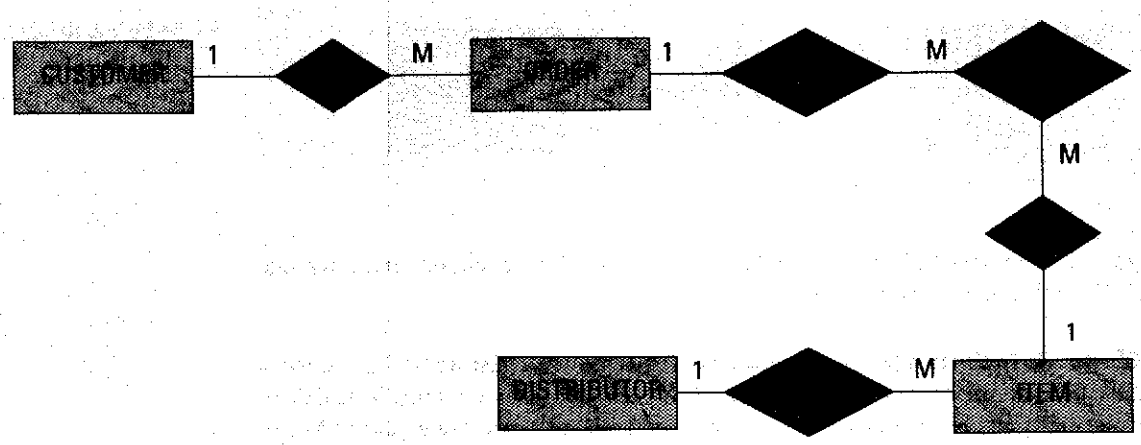



**FIGURE C.7**
A Composite Entity

Entities that exist to represent the relationship between two other entities are known as ***composite entities.*** The preceding example needs another entity that breaks up the many-to-many relationship between *ORDER* and *ITEM*. Figure C.7 displays the new relationship.

Creating a composite entity called *LINE ITEM* (think of it as a line item on an invoice slip) breaks up the many-to-many relationship between *ORDER* and *ITEM*, which then eliminates redundancy and other anomalies when deleting or updating information. Using the Chen model, composite entities are documented with a combination of a rectangle and a diamond.

Given the new ERD in Figure C.7, each *ORDER* can contain many *LINE ITEM*(s), but a *LINE ITEM* can belong to only one *ORDER*. As a result, the relationship between *ORDER* and *LINE ITEM* is one-to-many (one order has one or more line items), and the relationship between *LINE ITEM* and *ITEM* is one-to-many (one item can be in many line items). The composite entity has removed the original many-to-many relationship and turned it into two one-to-many relationships.

## Relationship Cardinality

***Cardinality*** expresses the specific number of instances in an entity. In the Chen model, the cardinality is indicated by placing numbers beside the entities in the format of (x, y). The first number in the cardinality represents the minimum value and the second number is for the maximum value.

Mega-Video can store data about a *CUSTOMER* in its database before the *CUSTOMER* places an *ORDER*. An instance of the *CUSTOMER* entity does not have to be related to any instance of the *ORDER* entity, meaning there is an optional cardinality.

However, the reverse is not true for the Mega-Video database; an *ORDER must* be related to a *CUSTOMER*. Without a *CUSTOMER*, an *ORDER* cannot exist. An instance of the *CUSTOMER* entity can be related to zero, one, or more *ORDER*(s) using the cardinality notation (0,N). An instance of the *ORDER* entity must be related to one and only one *CUSTOMER*, having a cardinality of (1,1). The relationship between an instance of *ORDER* and *CUSTOMER* is a mandatory relationship. Figure C.8 illustrates these cardinalities.

**FIGURE C.8**
Example of Cardinalities

## RELATIONAL DATA MODEL AND THE DATABASE

**LO C.3: Explain the need for an entity-relationship diagram in a database management system.**

Once the ERD is completed, it can be translated from a conceptual logical model into the formal data model required by the DBMS. The relational data model is the result of the work of Edgar (E. F.) Codd, a mathematician. During the 1960s, Dr. Codd was working with existing data models when he noticed that data relationships were very inefficient. Using his experience and knowledge in mathematics, he created the relational data model. Most databases, such as Access 2010 and SQL Server 2010, are based on the relational data model.

### From Entities to Tables

In creating an ERD for the conceptual model, the focus was on identifying entities and attributes. For the logical relational data model, the attention is on tables and fields. Using the ERD, the entities become tables and attributes turn into fields. A ***table*** is composed of rows and columns that represent an entity. A ***field*** is a characteristic of a table. A ***record*** is a collection of related data elements. The columns in the table definition represent the field, and a row is a record.

At first glance, a table along with the fields and records looks much like information in a spreadsheet, such as that displayed in Figure C.9 of a *CUSTOMER* table.

**Fields** Figure C.9 has four fields, *Customer Number, First Name, Last Name,* and *Phone Number.* Two or more tables within the same relational data model may have fields with the same names, but a single table must have unique field names. Using the relational data model notation, the table names are capitalized (e.g., *CUSTOMER*) and all columns are in title case (e.g., Customer Number), as in:

CUSTOMER (Customer Number, First Name, Last Name, Phone Number)

**Records** A record in a table has the following properties:

- A table cannot have multivalued attributes (as mentioned previously); therefore, only one value is allowed at the intersection of a field and record.
- Each record needs to be unique; there are no duplicate records in a table.
- A record must have an entity identifier, or ***primary key,*** which is a field (or group of fields), that uniquely identifies a given record in a table.

**Primary Key** A primary key makes it possible to identify every record uniquely in a table. The primary key is important to retrieve data accurately from the database.

Using a *Customer Number* as a primary key states that no two customers will ever have the same number. The primary key will be used to identify records associated with it. For example, if someone was searching the Mega-Video database for all the *ITEMS* that a *CUSTOMER* with a *Customer Number* of "112299" bought, he would retrieve only those records and not those associated with another customer.

Along with being unique, a primary key must not contain the value *null.* Recall that null is a special value meaning unknown; however, it is not the same as a field being blank or

**CUSTOMER**

| Customer Number | First Name | Last Name | Phone Number |
|---|---|---|---|
| 0001 | Bill | Miller | 777-777-7777 |
| 0505 | Jane | Cook | 444-444-4444 |
| 1111 | Sam | Smith | 555-555-5555 |
| 1212 | John | Doe | 666-666-6666 |

**FIGURE C.9**
A Sample Customer Table

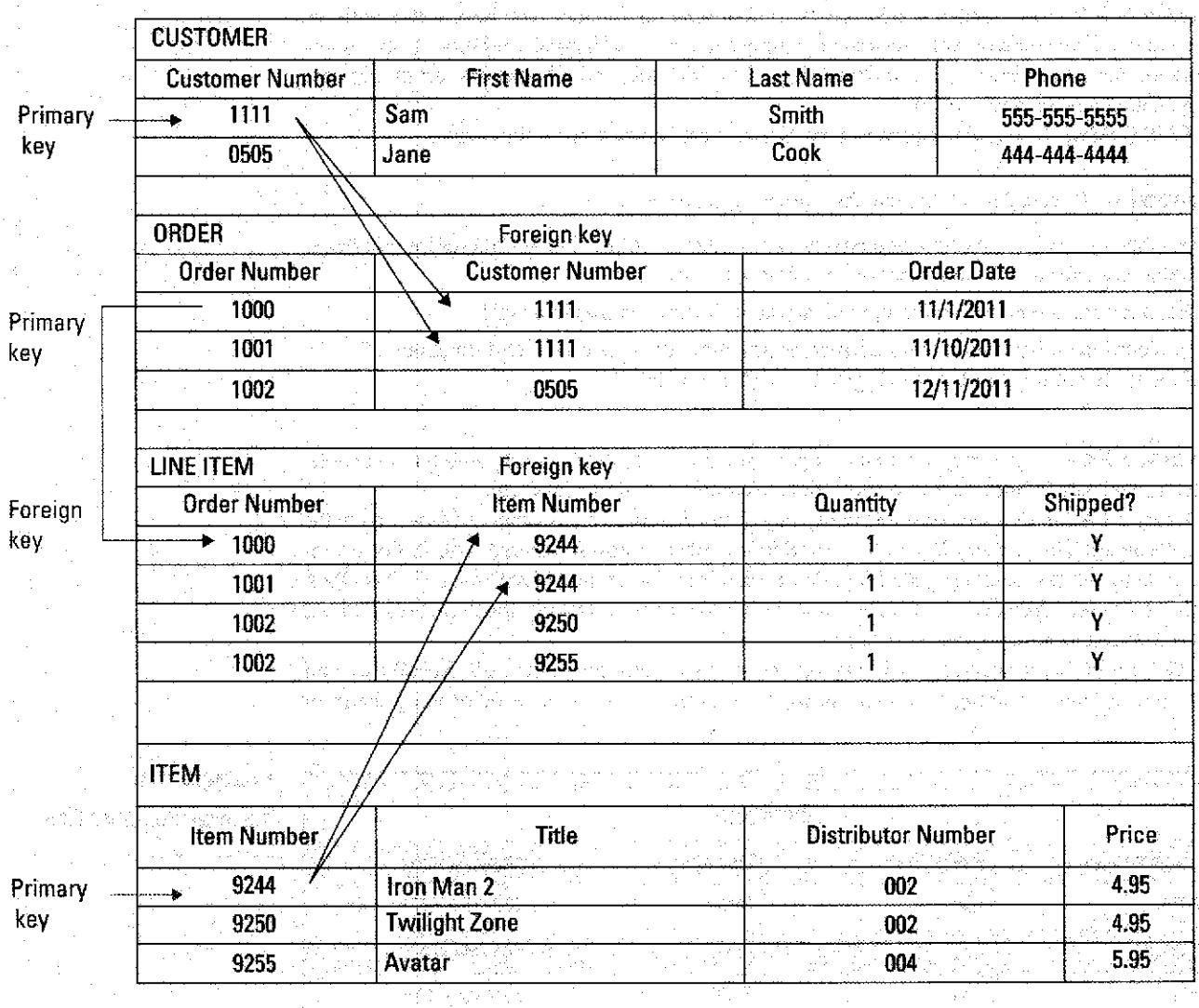
set to the value of zero. If one record has a null primary key, then the data structure is not in violation. But once a second null value for another record is introduced, the uniqueness of the primary key is lost. Therefore, nulls are forbidden when establishing primary keys.

The proper notation to use when documenting the primary key is to underline it, such as:
CUSTOMER (Customer Number, First Name, Last Name, Phone Number)

## Logically Relating Tables

Once the primary key has been defined, tables can be logically related. Each table in Figure C.10 is directly analogous to the entities of the same name in the Mega-Video ERD presented in Figure C.8, excluding the *DISTRIBUTOR*. The *CUSTOMER* table is identified by a Customer Number, a randomly generated unique primary key. The *ORDER* table is identified by an Order Number, another arbitrary unique primary key assigned by Mega-Video. The table *ORDER LINE* tells the company which *ITEM*(s) are part of which *ORDER*. This table requires a concatenated primary key (that is to say, joining two fields that act as one primary key) because multiple *ITEM*(s) can appear on multiple *ORDER*(s). The selection of this primary key, however, has more significance than simply identifying each record; it also represents a relationship among the *ORDER LINES*, the *ORDER* on

**FIGURE C.10**
Logically Relating Tables



which they appear, and the *ITEM*(s) being ordered. The primary key for the *ITEM* table is identified by the *Item Number.*

The *Item Number* field in the *ORDER LINE* table is the same as the primary key in the *ITEM* table. This indicates a one-to-many relationship between the two tables. Similarly, there is also a one-to-many relationship between the *ORDER* and *ORDER LINE* tables because the *Order Number* column in the *ORDER LINE* table is the same as the primary key of the *ORDER* table.

When a table contains a field that is the same as the primary key of another table, it is called a foreign key. A ***foreign key*** is a primary key of one table that appears as an attribute in another table and acts to provide a logical relationship between the two tables. The matching of foreign keys to primary keys represents data relationships in a relational database.

Foreign keys may be a part of a concatenated primary key, as is the case in the *LINE ITEM* table in Figure C.10. By concatenating, or combining, both *Order Number* and *Item Number* in the *LINE ITEM* table as foreign keys, they then become primary keys. However, most foreign keys are not part of the table's primary key. Consider the relation between Mega-Video's *CUSTOMER* and *ORDER* in Figure C.10. The *Customer Number* field in the *ORDER* table is a foreign key that matches the primary key of the *CUSTOMER* table. It represents the one-to-many relationship between *CUSTOMER* and *ORDER.* However, the *Customer Number* is not part of the primary key of the *ORDER* table; it is simply used to create a relationship between the two tables, *CUSTOMER* and *ORDER.*

A relational database uses the relationships indicated by matching data between primary and foreign keys. Assume that a Mega-Video employee wanted to see what *Titles* had been ordered with *Order Number* 1002. First, the database identifies the records in the *LINE ITEM* table that contain an *Order Number* of 1002. Then, it matches them to the *Item Number*(s) in the *ITEM* table. The results are those that match the records from each table.

## KEY TERMS

Attributes, C.2
Business rule, C.2
Cardinality, C.3
Composite entity, C.6
Data model, C.2
Database, C.1
Database management system (DBMS), C.1
Entity, C.2
Entity-relationship diagram (ERD), C.2
Field, C.7
Foreign key, C.9
Many-to-many relationship (M:N), C.5
Multivalued attribute, C.2
Null-valued attribute, C.3
One-to-many relationship (1:M), C.5
One-to-one relationship (1:1), C.4
Primary key, C.7
Record, C.7
Relational database model, C.2
Single-valued attribute, C.2
Table, C.7

## APPLY YOUR KNOWLEDGE

### 1. SportTech Events

SportTech Events puts on athletic events for local high school athletes. The company needs a database designed to keep track of the sponsor for the event and where the event is located. Each event needs a description, date, and cost. Separate costs are negotiated for each event. The company would also like to have a list of potential sponsors that includes each sponsor's contact information, such as the name, phone number, and address.

Each event will have a single sponsor, but a particular sponsor may sponsor more than one event. Each location will need an ID, contact person, and phone number. A particular event will use only one location, but a location may be used for multiple events. SportTech asks you to create an ERD from the information described here.

### 2. Course and Student Schedules

Paul Bauer, the chair for the information technology (IT) department at the University of Denver, needs to create a database to keep track of all the courses the department offers. In addition to the course information, Bauer would like the database to include each instructor's basic contact information, such as ID number, name, office location, and phone number. Currently, the department has nine instructors (seven full-time faculty and two adjuncts). For each course, Bauer would like to keep track of the course ID, title, and number of credit hours. When courses are offered, the section of the course receives an ID number, and with that number, the department keeps track of which instructor is teaching the course. There is only one instructor per course.

Finally, Bauer needs to be able to keep track of the IT students and to know which courses each student has taken. The information he would like to know about each student includes ID number, name, and phone number. He also needs to know what grade the student receives in each course.

He has asked you to create an ERD from the information described here using the Chen model.

### 3. Foothills Athletics

Foothills Athletics is an athletic facility offering services in Highlands Ranch, Colorado. All property owners living in Highlands Ranch are members of the Highlands Ranch Community Association (HRCA), which has partnered with Foothills Athletics to provide recreation facilities for its residents. Foothills Athletics has been using a spreadsheet to keep track of its personnel, facilities, equipment, and the HRCA members. The spreadsheet has created many redundancies along with several anomalies in adding, modifying, and deleting information. One of the HRCA members has suggested that the athletic facility should create a database to improve data collection that will also remove many of the difficulties that the spreadsheet is creating.

Foothills Athletics primary business operations are based on the following:

- **Personnel:** Foothills Athletics has a number of employees, primarily fitness instructors and administrative personnel. Records are kept on each employee, detailing employee name, address, phone number, date of hire, position, and status as either a current or former employee. Employees are assigned a unique four-digit employee ID number when they are hired.
- **Members:** When joining Foothills Athletics, HRCA members are assigned a unique four-digit member ID number. This information along with their name, address, phone number, gender, birth date, and date of membership are recorded. At the time of enrollment, each member decides on one of three available membership types along with a fixed membership fee: Platinum ($400), Gold ($300), and Silver ($200). This is a one-time fee that establishes a lifetime membership.
- **Facilities and equipment:** Foothills Athletics has a variety of facilities and equipment choices. Each facility has a unique room number and a size limitation associated with it. Some of the rooms contain a variety of exercise equipment; all have a serial number that is used for inventory and maintenance purposes. In addition, for each piece of equipment, the purchase date and the date of its last maintenance are recorded. Each piece of equipment belongs to a specific equipment type, such as elliptical machine, and is assigned a unique three-digit identification number. The description, the manufacturer's model number, and the recommended maintenance interval for that model of equipment are also kept on file. Each equipment type is associated with a single manufacturer that is referenced by a unique two-digit manufacturer ID number.

You have been hired to assist Foothills Athletics to create an ERD from the information described here using the Chen model.